Санкт–Петербургский государственный университет

ПОЛОЗ Алексей Евгеньевич

Выпускная квалификационная работа

Методы анализа пользовательских дискуссий в сети Web 2.0 на примере Telegram

Уровень образования: бакалавриат
Направление 02.03.02 «Фундаментальная информатика и информационные технологии»
Основная образовательная программа СВ.5003.2016 «Программирование и информационные технологии»
Профиль «Автоматизация научных исследований»

Научный руководитель:
доцент, кафедра технологии программирования,
к.т.н. Блеканов Иван Станиславович

Рецензент:
доцент, кафедра компьютерных технологий и систем, к.ф. - м.н. Погожев Сергей Владимирович

Санкт-Петербург
2020 г.

Содержание

Введение

Термин Web 2.0 впервые был использован в статье «Tim O'Reilly — What Is Web 2.0» от 30 сентября 2005 года. Под этим термином подразумевалась общая тенденция развития интернет-сообщества. Суть отражается комплексным подходом к организации, реализации и поддержке Web-ресурсов. Сейчас под сетями Web 2.0 понимают сети, которые путём учёта сетевых взаимодействий становятся тем полнее, чем больше людей ими пользуются. Таким образом особенностью Web 2.0 является принцип привлечения пользователей к наполнению.

К сетям Web 2.0 в первую очередь относятся социальные сети, системы мгновенного обмена сообщениями (далее мессенджеры) и форумы. Но подходы Web 2.0 распространены также и в сетях, не ставящих своей целью обмен пользовательским контентом, например, онлайн-магазины и новостные сервисы. В данных сетях используют такие подходы Web 2.0, как возможность оставлять комментарии, писать свои посты, заполнять профиль, подписываться на каналы, отправлять сообщения, составлять собственную новостную ленту и т.д.

Наибольшее распространение получили социальные сети. Они различаются типом контента и способами взаимодействия пользователей. Но основным принципом остаётся наполнение пользовательским контентом. Такой подход получил название User Generated Content (UGC). Пользователи помимо генерации контента могут выражать свои эмоции (оставлять реакции под постами), присоединяться к сообществам (подписываясь на каналы или вступая в чаты), выражать собственное мнение (используя упоминания, цитирование, репосты и комментарии).

Все вышеперечисленные способы взаимодействия поддаются анализу. Пользователи в процессе взаимодействия друг с другом создают дискуссию, определяют её тональность, выражают проблематику, связывают с другими темами. В таком обсуждении уже можно выделить лидеров мнений, стороны обсуждения, объект дискуссии, общую заинтересованность или остроту проблемы. Участники дискуссии могут неосознанно выражать собственную точку зрения, либо наоборот: манипулировать мнением других людей, пре-

следуя определённые цели и мотивы.

Проанализировав такую дискуссию можно получить важную информацию о текущем положении дел или общее отношение пользователей к конкретной тематике, продукту, событию или идее. Более того, разнообразие подходов для взаимодействия даёт возможность составить географическую карту настроения пользователей. К таким подходам относятся гео-чаты (чаты привязанные к геопозиции), локальные чаты (чаты сообществ и муниципалитетов) и геометки конкретных постов.

Получение списка пользователей сообществ даёт возможность найти пересечение интересов пользователей, скрытые сообщества и источники мнений, перетекающих из одного обсуждения в другое. Анализ активности и взаимодействия с другими пользователями может указать на транслирование определённой информации, преследуя личные мотивы, распространение пропаганды или «вбросов» — заведомо ложной информации, в целях создания общей паники, распространению продукта или в иных целях.

Граф связей сообщений позволяет выделить объект обсуждения, определить взаимосвязанные дискуссии. В определённых случаях «боль» пользователя и источник его утверждений можно определить лишь выделив частную дискуссию или «идущую параллельно», т.е. дискуссию по другой тематике, имеющей косвенную связь с текущей.

Актуальность

Рынок сетей Web 2.0 уже занимает неотъемлемую часть жизни пользователей сети Интернет и продолжает увеличиваться. По данным аналитического сервиса SimilarWeb на 1 апреля 2020 года социальная сеть Facebook является самым посещаемым сайтом в категории «Social Networks and Online Communities», занимая 3 место в мире в общем рейтинге веб-сервисов, уступая поисковой системе Google и видеохостингу YouTube [1].

Важно отметить активность Web 2.0. Только с 2013 по 2019 год было зарегистрировано более 2 миллиардов уникальных пользователей в социальных сетях [2]. А количество активных пользователей Facebook насчитывает более 2,6 миллиарда ежемесячно по данным на первый квартал 2020 года [3].

Не менее важным игроком в Web 2.0 являются мессенджеры. В последнее время они активно набирают популярность, так например, Facebook принял решение разделить социальную сеть и мессенджер. Общение внутри Facebook происходит с помощью специального мессенджера Messenger, который занимает второе место в мире в категории «Мессенджеры». Первое место занял WhatsApp [4].

Мессенджеры к настоящему моменту уже занимают первое место среди приложений по среднему дневному охвату на мобильных устройствах, социальные сети на втором месте. При этом по количеству времени, проведённому внутри приложения, социальные сети лидируют со средним показателем 52 минуты на пользователя в день [5].

Инструменты

Многие современные мессенджеры ставят целью создание целой экосистемы благодаря реализации базового функционала. Так например, мессенджер WeChat, лидер китайского рынка, позволяет проводить денежные операции, оформлять ряд документов, обрабатывать фотографии, подтверждать личность при пересечении границы между материковым Китаем и Гонконгом вместо паспорта, знакомиться со случайными людьми. Также в мессенджере доступен сервис машинного перевода. Количество пользователей WeChat на первый квартал 2020 года составляет 1,2 миллиарда [6].

Одним из наиболее распространённых мессенджеров с развитой архитектурой для интеграции таких возможностей является Telegram. Прирост пользователей Telegram за последний год составил 100 миллионов пользователей, с общим числом в 400 миллионов (по данным на 24 апреля 2020 года) [7]. Уже в 2016 году пользователями ежедневно отправлялось 15 миллиардов сообщений [8].

С помощью специальных средств в Telegram возможно создавать группы, «супергруппы», каналы, геочаты, есть возможность привязывать чаты к каналам, реализована сущность «бот», возможно оставлять сообщения с геолокацией, подписями, создавать собственные реакции к постам, оставлять комментарии, обращаться к конкретным пользователям, ссылаться на сообщения или пересылать их, реализована система распределения прав в каналах и группах.

Особенно интересны возможности API, именно это подтолкнуло данный мессенджер к интенсивному развитию, так как простой интерфейс с широким спектром возможностей позволяет создавать боты всевозможных форматов. Для упрощения взаимодействия с пользователями компании используют ботов в Telegram, что позволяет заменить собственные приложения и сервисы. Показательным примером такой компании является проект «Карта Города», поддерживаемый Администрацией Санкт-Петербурга. Внутри бота можно оплачивать проезд в метро и другом общественном транспорте, получать скидки и аналитику по проезду.

Именно благодаря глубокой интеграции ботов в процесс общения, Telegram

открывает широкий выбор форматов взаимодействия пользователей. Это позволяет сложить более целостную картину при анализе дискуссий. Так например, возможность оставлять реакции реализуется с помощью специального бота, в котором возможно создавать уникальные реакции под каждое сообщение. Подобный подход внедрил Facebook в 2016 году, добавив помимо реакции «Нравится» ещё 5 новых реакций, которые зависят от содержимого поста.

Важным фактором при выборе сети для анализа дискуссий была открытость сети для получения пользовательского контента. Telegram позволяет реализовывать ботов, которые могут извлекать посты за весь период существования каналов и чатов, производить гибкий поиск по запросу, а также мессенджер не имеет жёстких ограничений API. Аналитики отмечают высокую активность русскоговорящей аудитории. Вокруг Telegram сформировалось целое сообщество. Чаты с открытым обсуждением достигают 100 000 пользователей.

По данным показателям Telegram соответствует всем необходимым требованиям для сбора данных и проведения анализа дискуссий. Сообщество активно поддерживает развитие мессенджера, а уже существующие возможности позволяют анализировать сети Web 2.0. По сравнению с мессенджерами в социальных сетях не так развита инфраструктура для массового общения большого количества пользователей в чатах.

Цель работы

Целью данной работы является исследование методов и инструментов для анализа пользовательских дискуссий в сетях Web 2.0 на примере Telegram и выявление скрытой семантически значимой информации о предмете дискуссии и пользователях, влияющих на неё.

Решение обозначенной глобальной цели будет способствовать выявлению кибербулинга, мошенничества, реакционно настроенных сообществ и террористических группировок.

Анализ мессенджеров важен для бизнеса, так как там общается целевая аудитория — потенциальные и действующие клиенты. Направление мессенджер-маркетинг это важная часть общего интернет-маркетинга, так как помогает бизнесу решать задачи выстраивания коммуникации (техническая поддержка, реакция на негативные отзывы, сбор обратной связи), формирования лояльности аудитории (улучшение имиджа и узнаваемости, построение сообщества вокруг бренда) и задачи доставки контента (информирование о новых продуктах, акциях и скидках, распродажах). Решение этих задач важно для маркетинговых агентств, так как позволяют увеличивать продажи и снижать расходы на реализацию, поддержку и анализ.

Постановка задачи

Результатом данной работы станет веб-сервис, позволяющий в режиме онлайн проводить анализ и получать визуализацию по выбранным тематикам. Для достижения поставленной цели необходимо выполнить следующие шаги:

- Провести обзор существующих инструментов по анализу
- Спроектировать архитектуру программного комплекса для Telegram на предмет выявления и анализа дискуссий
- Реализовать метод для выгрузки пользовательского контента из сети Telegram
- Провести предварительную обработку данных и добиться наилучших показателей модели
- Исследовать и сравнить методы анализа дискуссий
- Визуализировать полученные результаты
- Протестировать и апробировать на реальном кейсе

Практическая значимость

Задача анализа дискуссий в сетях достаточно важна, так как является междисциплинарной, решаются как социально-экономические задачи (таргетинг, оценка пользовательского настроения), политические задачи (геополитический анализ настроений населения, выявление текущих и появляющихся проблем), так и задачи менеджмента (анализ актуальности и определение ценностей пользователей). Всё это объединяет специальная наука – Social Network Analysis (SNA).

В современном мире информация, полученная из анализа пользовательской активности, определяет векторы развития компаний, а также приводит к возникновению новых рынков и ниш для бизнеса. Во многих современных стартапах используется подход с поиском «болей» пользователей, выдвижением гипотез и их тестированием. Весь процесс завязан на глубокой интеграции в пользовательское взаимодействие. Очень важно понимать что является основополагающим источником проблем и каковы его последствия. Касается это не только бизнеса, но и играет огромную роль в политике, экономике, бренд-менеджменте и ряде других сфер. Компании проводят оценку рисков на основе реакций на определённые продукты, решения, события. Знание географических и временных рамок очагов проблем могут сыграть ключевую роль в становлении будущего бизнеса и государств. Понимание подхода к решению проблем может качественно изменить оценку решений, вплоть до определения будущих тенденций и предотвращения проблем в корне.

Глава 1. Обзор существующих решений

1.1 Технологический обзор

Уже существуют SMM платформы, которые позволяют анализировать тональность мнений покупателей для оценки имиджа бренда. Одним из таких представителей является YouScan — международная компания, разрабатывающая решения для мониторинга и аналитики социальных медиа с 2009 года. Основным функционалом платформы является визуальная аналитика (анализ пользовательских изображений: определение бренда, типа изображения, объектов, персон и рода деятельности), распознание трендов (график упоминаний, анализ вовлечённости и увлечённости, количество уникальных авторов, оценка пола аудитории, основной сети обсуждения и геопозиции) и распределение тональности по времени [9]. Данное решение является платным и проводит анализ в социальных сетях, блогах и онлайн медиа, на форумах и сайтах отзывов, не затрагивая групповые чаты в мессенджерах. YouScan не предоставляет анализа конкретных дискуссий и взаимосвязей между ними.

Поисковые системы также предоставляют аналитику по поисковым запросам, веб-приложениям и брендам. Аналитический сервис Google Trends позволяет получать графики активности пользователей по запросу в заданный промежуток времени на определённой территории. Запросы оцениваются в условных единицах, доступна возможность сравнения различных запросов. Яндекс имеет несколько сервисов для подобной аналитики. Яндекс Wordstat позволяет отслеживать поисковые тренды, показывает связанные запросы и возможно близкие тематики. Яндекс.Радар формирует топ интернет-проектов, поисковых систем и технологий в выбранной категории на заданном промежутке времени, с возможностью сортировки по устройствам, размеру дохода, возрасту и полу. Стоит отметить, что данные сервисы приводят аналитику по поисковым запросам, поэтому такой анализ не является репрезентативным для анализа пользовательских дискуссий.

1.2 Обзор существующих методов анализа

Исследователи в количественном анализе опираются на социологическую теорию сетей Web 2.0, математическим базисом которой является теория графов. В такой системе узлами являются пользователи, а рёбрами графа – связь пользователей: обращения, упоминания, ответы. Благодаря этому становится возможным анализ паттернов распространения или анализ неключевых участников дискуссии, для этого используют термин «показатель центральности узла». Количественные методы включают в себя статистику, регрессионный анализ, работу с частотными словарями, временные графы дискуссий, алгоритмы фильтрации сообщений и выделения ключевых слов [10].

К качественным методам относят ситуационный анализ, интерпретативное чтение, структурно-функциональный анализ веб-графов, сопоставительный и сравнительный анализ различных элементов полученных данных. В частном случае качественный анализ сообщений заключается в том, что после выделения обобщённых тем и связанных с ними наборов слов происходит сортировка этих топиков на относящиеся к области исследования и не относящиеся [10].

Для моделирования топиков исследователи используют алгоритмы Latent Dirichlet Allocation (LDA) и Latent Semantic Analysis (LSA) [11]. Также для анализа коротких документов, таких как сообщения в переписке пользователей используется Biterm Topic Model (BTM) [12]. В данной работе будут рассмотрены методы анализа тональности сообщений, т.е. сентимент-анализ, обработки модели языка и моделирования тематик (топиков).

Глава 2. Разработка программного комплекса

Программная обработка сложных потоков информации включает два направления:

- **Накопление информации** - создание и ведение баз данных специализированной информации, являющейся частью интегрированной базы;
- **Обобщение информации** - построение пространственно-информационных запросов к интегрированной базе, результаты которых принимают вид постоянно обновлённых карт по различным направлениям деятельности центра и позволяют решать аналитические задачи.

2.1 Объекты взаимодействия

Первоначально необходимо определить какие форматы взаимодействия подразумевают сети Web 2.0. В процессе анализа были выделены следующие объекты взаимодействия:

- **Каналы** – различные публичные страницы, группы с публикациями, страницы авторов. Отличительной особенностью является публикация постов ограниченным числом пользователей, являющимися владельцами или модераторами данного канала. Общедоступное обсуждение может быть в комментариях к основному посту. Примером каналов можно назвать «Сообщества» в социальной сети ВКонтакте, «Группы» в социальной сети Facebook, «Каналы» на видеохостинге YouTube, «Каналы» в мессенджере Telegram.
- **Чаты** – различные обсуждения и групповые переписки. Отличительной особенностью является свободное ведение дискуссии всеми участниками чата. Примером чатов можно назвать «Чаты» в социальной сети ВКонтакте, «Группы» или «Супергруппы» в мессенджере Telegram.
- **Посты** – публикации различных форматов. Это могут быть «Новости» в социальной сети ВКонтакте, представляющие собой текстовую публикацию с прикреплением всевозможных форматов документов, «Посты»

фото или видеоформата в социальной сети Instagram, «Сообщения» в мессенджере Telegram.

- **Реакции** – различные подходы, с помощью которых пользователи могут показать своё отношение к публикуемому контенту. Отличительной особенностью является выражение эмоций. Примером реакций можно назвать «Лайки» в социальной сети ВКонтакте, «Нравится», «Ух ты!», «Супер», «Сочувствую», «Возмутительно», «Ха-ха», «Мы вместе» в социальной сети Facebook, «Нравится»и «Не нравится» на видеохостинге YouTube. В Telegram нет базовой возможности оставлять подобные реакции, но есть возможность прикреплять кнопки к сообщениям, таким образом благодаря этому функционалу любой пользователь может создать свою собственную реакцию. К реакциям также стоит относить другие форматы взаимодействия пользователей, например: «просмотры», «комментарии» и «репосты».

- **Профили** – личные аккаунты пользователей. В личном аккаунте пользователь может указать имя, фамилию, описание, геопозицию, имя пользователя (логин), личную информацию, оставить ссылки на аккаунты в других сетях. Также в некоторых сетях можно увидеть подписки (на кого подписан пользователь) и подписчиков (кто подписан на пользователя), тем самым определив взаимосвязь между несколькими участниками дискуссии.

В текущей работе будут рассматриваться исключительно посты в чатах, т.к. публикации в каналах Telegram не позволяют сформировать пользовательские дискуссии. Также в профилях нельзя получить информацию о взаимосвязи с другими участниками сети, единственной ценностью, которую можно извлечь из профилей Telegram, является личная информация, она может указывать на геопозицию или профили в других сетях.

2.2 Проектирование архитектуры

Проект состоит из нескольких модулей полного цикла – от сбора пользовательского контента до визуализации аналитических выводов.

Для реализации был выбран следующий стек технологий:

- DevOps & Серверная составляющая
  - **Docker** – программное обеспечение для контейнеризации проекта и автоматизации развёртывания на кластере. В данном проекте используется для изолированной разработки в разных средах, а также для быстрого запуска при внесении больших изменений в архитектуру.
  - **NGINX** – веб-сервер. Был выбран поскольку имеет достаточно простой функционал, включающий в себя всё необходимое для реализации данного проекта. NGINX имеет активную поддержку сообщества, глубокую интеграцию с другими продуктами, например: Docker и Let's Encrypt, а также хорошо документирован.
  - **Let's Encrypt** – real-time сервис центра сертификации, позволяющий обеспечивать соединение шифрованием HTTPS. Выбран исходя из того, что является самым распространённым некоммерческим решением.
- Back-end
  - **Flask** (на основе **Python**) – микрофреймворк для создания веб-приложений. Является небольшим по размеру решением, позволяющим гибко реализовывать функции API. Был выбран фреймворк на языке Python с целью производить аналитические вычисления.
  - **MongoDB** – документоориентированная система управления базами данных (далее СУБД). Имеет достаточно высокую скорость для нереляционных СУБД.
  - **Socket.IO** – библиотека для обмена данными (сокетами) в режиме реального времени.

- Front-end
  - **React** (на основе **JavaScript**) – JavaScript-библиотека для разработки пользовательских интерфейсов и каркаса веб-приложений. Является распространённым решением для реализации компонентного подхода.
  - **Redux** – JavaScript-библиотека, предназначенная для управления состоянием веб-приложения. Используется в связке с React.
  - **Bootstrap** – фреймворк для создания веб-интерфейса приложения. Реализует базовые элементы и адаптивность под различные устройства.

2.2.1 Стратегия обхода

Выгрузка публикаций происходит по заданной области исследования. Каждая область исследования состоит из ключевых слов, которые определяют эксперты. Например, для области исследования «криптовалюта» - ключевыми словами будут являться «токен», «блокчейн», «распределённый реестр» и т.д. По каждому из ключевых слов происходит поиск средствами Telegram в отобранных чатах. Стандартный поиск Telegram позволяет находить поисковые запросы в разных формах и склонениях русского языка.

2.2.2 Обработка данных

На этапе обработки данных происходит приведение естественных текстов в более формальный, структурированный вид, удобный для анализа. Обработка состоит из следующих этапов:

1. **Лемматизация** – все слова приводятся в начальную форму.
2. **Формирование «мешка слов»** – каждая публикация разбивается на список слов (шинглов) и формируется словарь, по которому далее будет происходить индексирование.
3. **Фильтрация по стоп-словам** – определяются и исключаются слова, не имеющие отношения к объекту анализа.

4. **Фильтрация по части речи** – исключаются числительные, союзы, предлоги и т.д.

5. **Фильтрация по частотности** – отсекаются слишком редкие и слишком частые слова (шинглы).

6. **Фильтрация по размеру** – после всех преобразований отсеиваются слишком короткие, несущественные сообщения.

7. **Сентимент-анализ** – определяется тональность текста, далее отсеиваются публикации с положительной и нейтральной тональностью.

2.2.3 Формирование тематик

Для задачи выявления топиков используются такие методы тематического моделирования, как Latent Dirichlet Allocation (LDA). Тема формируется из списка слов и коэффициентов вхождения этого слова в тематику. Стоит отметить, что любой документ (публикация) с какой-то долей вероятностью соотносится с каждой из выявленных тематик. В данной работе для простоты анализа будем считать, что документ относится к наиболее вероятной тематике.

2.2.4 Визуализация

Проблемные темы отображаются на временной шкале, после чего определяется их актуальность в заданный промежуток времени. Результаты анализа обобщаются в таблице, именуемой «Тепловая карта». Каждая ячейка подсвечивается цветом – от тусклого к яркому. Благодаря такому подходу формируется понимание о том какие темы были актуальны, как долго и в какие смежные темы перешёл интерес.

2.3 Поиск релевантных источников

Для анализа дискуссий необходимо обрабатывать естественное общение пользователей, поэтому далее будут рассматриваться чаты. Из исследования были исключены каналы, где публикация происходит от ограниченного числа лиц и нет возможности обсуждения. В Telegram имеется возможность прикреплять чаты к каналам, что позволяет обсуждать опубликованную новость в специальном месте. Помимо этого, возможно прикреплять веб-страницу для комментирования конкретных записей, но данный формат не является популярным.

Для выгрузки сообщений отбираются чаты с относительно большим количеством пользователей (более 1000 человек) и достаточно активным обсуждением – последние 100 сообщений должны быть сделаны за последнюю неделю. Эти параметры были выявлены эмпирическим путём, для достижения большей точности в дальнейшем могут корректироваться.

Поиск производился с помощью специализированных чатов с подборками источников по категориям и на сайтах-агрегаторах. Также были проанализированы крупные сайты и группы в социальных сетях, которые развивают собственные сообщества. Такой подход позволил эффективно найти активные чаты с большой аудиторией, так как их развитием занимаются целые команды.

2.4 Извлечение пользовательского контента

Для сбора контента было разработано веб-приложение. В целях получения лучшего соединения с серверами Telegram приложение размещено на кластере DigitalOcean (Droplet) в городе Франкфурт-на-Майне.

Извлечения пользовательского контента реализуется Python-библиотекой «Telethon», основанной на протоколе MTProto. С помощью методов получения списка чатов определяются доступные чаты, отобранные на предыдущем этапе. Затем чаты фильтруются в соответствии с объектом исследования.

В Telegram сообщения различных форматов имеют разные типы данных. Это существенно отличается от подхода, используемого в API ВКонтакте, в котором посты представляют собой объект с текстовым полем, с возможностью прикрепления различных документов. В настоящей работе были отобраны сообщения только текстового формата.

2.5 Предварительная обработка данных

Лемматизация

На этапе лемматизации все выгруженные сообщения разбиваются на отдельные слова, каждое из которых приводится в начальную форму. После апробации нескольких подходов было решено использовать морфологический анализатор PyMorphy2. Данная библиотека позволяет получать наиболее точные результаты за наименьшее количество времени (по сравнению с PyMyStem3). В тестировании на коротких сообщениях PyMorphy2 позволял проводить анализ в среднем за 0.145 секунды, когда PyMyStem3 занимал 1.423 секунды.

Формирование «мешка слов»

Имея сообщения в виде наборов слов в начальной форме, можно составить словарь всех слов, которые используются в дискуссиях данной области исследования. Словарь этих слов называется «мешок слов», в данном случае он будет использоваться для дальнейшей фильтрации и селекции сообщений. Проиндексировав слова, мы можем векторизовать сообщения пользователей. Этот этап необходим для достижения лучшей точности в анализе «необразцовых» текстов, таких как сообщения пользователей.

Отбор стоп-слов

Так как анализ происходит в свободной дискуссии, необходимо заранее исключить из этой дискуссии темы, не относящиеся к области исследования. Стоп-слова зачастую отбираются опытным путём. В исследовании стало понятно, что в основном такие слова зависят от категории чата с общим обсуждением. Например, в локальных чатах чаще употребляются название местности или города, что может не относиться к теме дискуссии, но влиять на географическую активность.

Фильтрация по части речи

По аналогии с предыдущим этапом отсекаются слова, не влияющие на дискуссии и тональность высказываний. Такие слова имеют широкое употребление, но являются непоказательными.

Фильтрация по частотности

Для достижения наилучшего результата из словаря также убираются слишком частые и слишком редкие слова. Данный этап позволяет сбалансировать вхождение слов в тематику. Коэффициенты выставляются эмпирическим путём. В данном проекте были определено, что лучший результат достигается при отсечении слов, встречающихся реже 3 раз, такие слова не позволяют точно оценить влияние слова на тему. Поскольку сообщения зачастую являются короткими, то отсечение по верхней границе не используется. При эксперименте с отсечением наиболее частых слов (5% самых частовстречающихся), итоговая модель выдавала значительно худшие результаты.

Фильтрация по размеру

После преобразований словаря происходит векторизация сообщений. Короткие сообщения (документы) или пустые (состоящие из 0 слов) удаляются из анализа дискуссии.

Сентимент-анализ

Поскольку конечной целью исследования является выявление проблематики пользователей, определяется тональность текста. Определение происходит способом выделения ключевых слов, которые задают тональность всего сообщения. Далее отсеиваются публикации с положительной тональностью. Это необходимо для того чтобы не смешивать объекты дискуссии, имеющие положительный и отрицательный окрас, что достаточно сильно влияет на результаты исследования.

2.6 Методы анализа

Латентно-семантический анализ (LSA)

Первоначально эксперимент проводился с использованием LSA, это основополагающий метод тематического моделирования. Основная идея данного метода заключается в разложении матрицы «документы-термины» на отдельные матрицы «документы-темы» и «темы-термины». На практике данный метод оказался неподходящим, так как не учитывается количество вхождений и влияние каждого слова на документ.

Вероятностный латентно-семантического анализ (pLSA)

Следующим этапом был анализ дискуссий с помощью метода вероятностного латентно-семантического анализа. В данном методе строится вероятностная модель, метод основан на смешанном разложении. Общую формулу можно определить так:

$$P(D,W) = P(D)\sum_{T} P(T|D)P(W|T)$$

Или так:

$$P(D,W) = \sum_{T} P(T)P(D|T)P(W|T)$$

Где P(D) - вероятность встретить слово, P(T) - вероятность встретить тему, P(D|T) - вероятность документа в теме, P(W|T) - вероятность слова в теме.

Вид отличается от LSA, но в общем случае добавляется вероятностная трактовка [13]. Были получены лучшие результаты, чем в модели LSA, что привело к использованию следующего метода.

Латентное размещение Дирихле (LDA)

Для того, чтобы применить LDA к документам (сообщениям) необходимо преобразовать извлечённый контент в терм-документную матрицу. Это

матрица размером D x W, где D – количество документов, а W – размер словаря (корпуса) слов, полученного на этапе предварительной обработки. Значением матрицы в i-ой строке j-ом столбце является цифра, соответствующая количеству раз встречаемости j-го слова в i-м документе.

Как и в pLSA для данной терм-документной матрицы строятся матрицы распределения тем по текстам и матрица распределения слов по темам. Матрица распределения тем по текстам имеет размер D x T, где T – количество тем, это значение задаётся отдельно. В следующем пункте будут показаны результаты модели в зависимости от различных значений T. Матрица распределения слов по темам имеет размер T x W.

LDA, в отличие от предыдущего метода, делает предположения о случайном распределении векторов документов и векторов тематик [14]. Стоит отметить, что LDA является обобщением модели pLSA, которая эквивалентна LDA при априорном распределении Дирихле [15].

Данный метод показал наилучшие результаты и используется в итоговом решении. Все данные анализа сохраняются в СУБД и доступны для визуализации. Каждая тематика определяется списком слов. В следующем пункте отображены тематики, которые были выделены с помощью реализованного метода LDA.

2.7 Визуализация

Интерфейс

Интерфейс веб-приложения имеет 2 пункта: «Тепловая карта» и «Активность». В разделе «Тепловая карта» доступно поле для создания запросов и просмотра уже проанализированных (рис. 1).

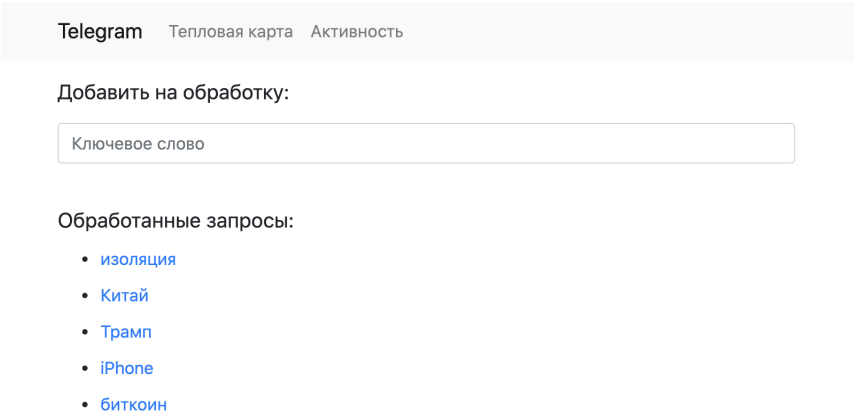

Рис. 1: Интерфейс запросов.

Запрос создаётся после указания пользователем ключевых слов, определяющих область исследования. Сбор данных и процесс анализа происходит на сервере и занимает до 5 минут, после завершения обработки ответ отправляется клиенту и отображается в поле «Обработанные запросы». Из-за задержек с обработкой было принято решение использовать сокеты, таким образом клиент не ожидает ответа с сервера. Все результаты запросов хранятся в истории и доступны для вторичного запроса без необходимости повторного анализа.

Тепловая карта

Для визуализации полученных тематик пользовательских дискуссий используется тепловая карта (рис. 2). Каждая ячейка отображает активность обсуждения определённой тематики во временной когорте. Активность рассчитывается в процентах относительно исторической активности тематики. Тема определяется списком слов с указанием коэффициента влияния.

Рис. 2: Тепловая карта по запросу «Яндекс».

Ниже приведена тепловая карта по запросу «Китай» (рис. 3). В первой строке наблюдается нарастание актуальности заданной дискуссии независимо от остальных. При этом в пике своей активности виден всплеск активности во второй дискуссии. Стоит отметить, что не имея информации о первой дискуссии было бы затруднительно предсказать такой рост.

На следующем рисунке можно отметить соответствие дискуссий заголовкам новостей (рис. 4). Ключевые слова достаточно точно описывают темы обсуждений и формируют понимание проблематики.

03.2019	04.2019	05.2019	06.2019	07.2019	08.2019	09.2019	10.2019	11.2019	12.2019	
0%	11%	11%	16%	21%	18%	20%	36%	46%	27%	0.020 • блокчейн 0.015 • проект 0.014 • финансовый 0.012 • технология 0.011 • китайский 0.010 • пользователь 0.010 • помощь 0.010 • пользоваться 0.010 • компания 0.009 • инвестор
3%	0%	3%	2%	13%	10%	10%	5%	100%	0%	0.038 • услуга 0.034 • европа 0.030 • украина 0.028 • перевод 0.026 • кэш 0.026 • позволять 0.025 • киев 0.025 • перестановка 0.025 • уточнять 0.023 • майами

Рис. 3: Нарастание интереса пользователей.

03.2019	04.2019	05.2019	06.2019	07.2019	08.2019	09.2019	10.2019	11.2019	
0%	0%	11%	11%	44%	33%	78%	78%	22%	0.014 • часть 0.014 • говорить 0.013 • доллар 0.013 • время 0.013 • масштаб 0.013 • курс 0.012 • управлять 0.012 • рынок 0.012 • китай 0.012 • миллиард
4%	12%	4%	8%	32%	12%	24%	12%	28%	0.051 • дональд 0.025 • делать 0.021 • сказать 0.020 • новый 0.019 • знать 0.016 • дом 0.016 • понимать 0.015 • блокчейн 0.014 • открытый 0.014 • обама
9%	0%	6%	6%	18%	3%	6%	12%	6%	0.035 • сша 0.030 • путин 0.025 • решить 0.024 • страна 0.023 • россия 0.016 • президент 0.014 • человек 0.013 • мир 0.012 • закрыть 0.011 • мировой

Рис. 4: Тепловая карта по запросу «Трамп».

График активности

Для отображения общей тенденции и интереса пользователей был реализован график активности обсуждения (рис. 5). Таким образом становится возможным анализ обсуждаемости заданной конкретной дискуссии. Для бизнеса такая аналитика позволяет узнать актуальность проблемы и отметить её тенденции на разных этапах.

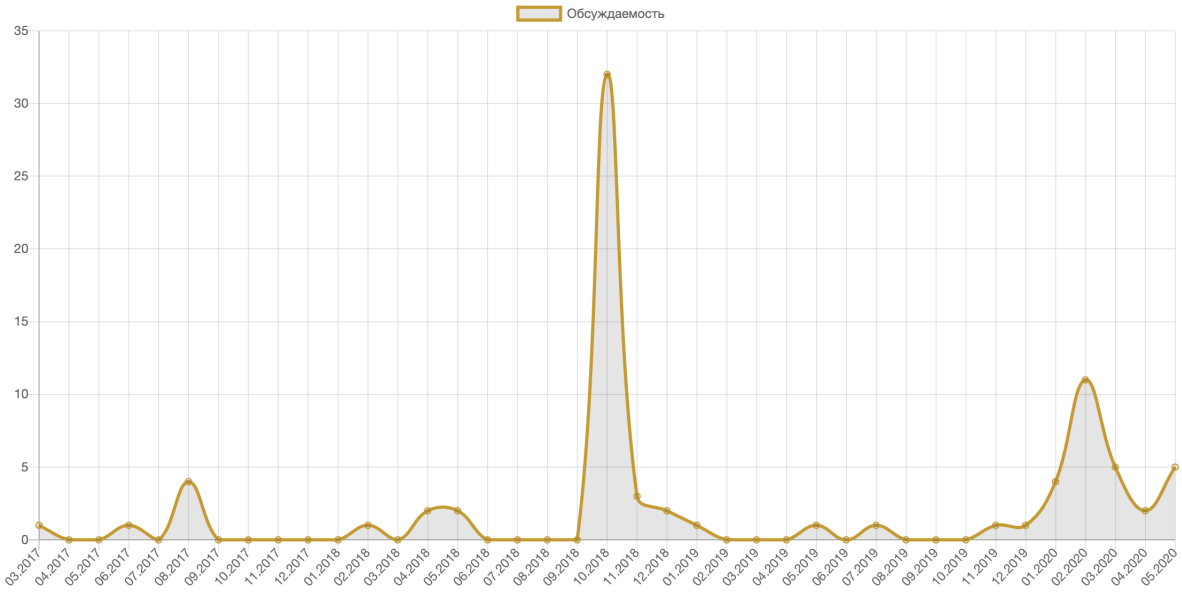

Рис. 5: График активности по запросу «Керчь».

2.8 Апробация

Запросы были апробированы на примере публичных личностей (Павел Дуров, Дональд Трамп), брендах (Яндекс, iPhone) и новостях (Керчь, самоизоляция). Зачастую на точность выделения ключевых слов в каждой теме влиял выбор их количества. Наиболее обсуждаемые тематики в некоторых случаях делились на две похожие, с одинаковыми показателями обсуждаемости.

Для подтверждения активности обсуждения дискуссий был использован аналитический сервис Google Trends (рис. 6). Общая тенденция совпадает (рис. 5), за исключением периода с декабря 2019 года по март 2020 года. Проанализировав в чём расхождение, было выявлено, что общий тренд в сервисе Google следует за активностью новостных источников, в то время как обсуждения в чатах могут зарождаться из близких тематик и обсуждений, это подтверждается анализом дискуссий.

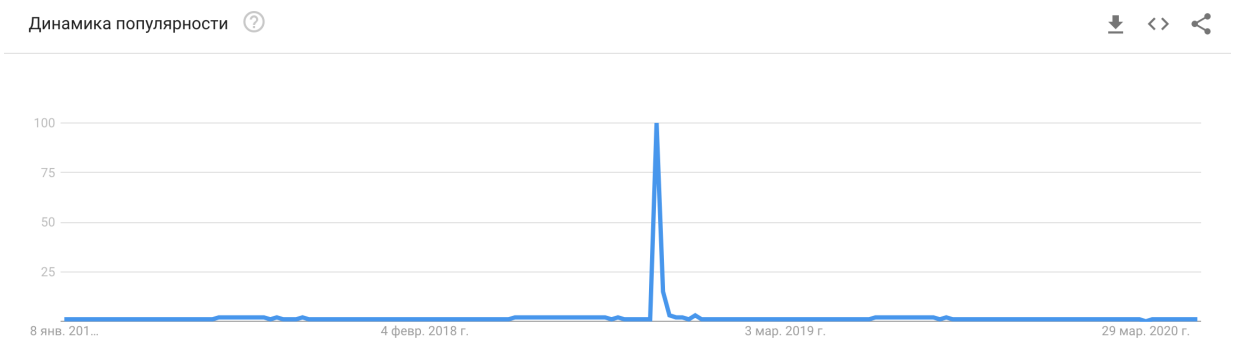

Рис. 6: Динамика популярности Google Trends по запросу «Керчь».

Рассмотрим кейс по запросу «изоляция». Обработка запроса заняла 74 секунды, из которых 14 – это сбор данных из Telegram и 55 секунд – предварительная обработка данных. После обработки запроса стала доступна тепловая карта (рис. 7). Проанализировав выделенные темы, видно, что слова в первой строке описывают тему «изоляция» в смысле «самоизоляция», на это указывают слова «карантин», «мир», «условие» и «жизнь». Вторая строка описывает закон об изолированном интернете в России, на это указывают слова «система», «россия», «интернет», «связь», «блокировка» и «закон». Для того чтобы избежать подобных пересечений, необходимо указывать больше ключевых слов, по которым происходит поиск. Также стоит

указывать стоп-слова, не влияющие на объект дискуссии и явно описывающие другую тематику. Частным случаем является пересечение тематик, так например в марте-апреле 2020 года эти тематики пересекаются, поскольку речь идёт об изоляции страны (закрытия границ) как временное решение на период самоизоляции (карантина).

Рис. 7: Тепловая карта по запросу «изоляция».

Вывод

Для получения наилучших результатов, необходимо указывать несколько ключевых слов, которые чаще всего могут встречаться в общении в общих чатах. Не менее важным является указание стоп-слов и регулирование количества тематик.

Заключение

Результаты работы

Было разработано решение для анализа пользовательских дискуссий в сетях Web 2.0 на примере Telegram. Код размещён в GitHub репозитории: https://github.com/kosyachniy/tg. В процессе работы был проведён обзор существующих инструментов анализа активности пользователей в UGC-сетях, спроектирована архитектура программного комплекса и реализован каркас веб-приложения. После были реализованы методы для извлечения пользовательского контента из сети Telegram и предварительной обработки данных. Проведено исследование работы методов LSA, pLSA и LDA. Затем по результатам модели построена визуализация.

Итоговая система была протестирована на объектах анализа, представляющих бренды, публичных личностей и новости. Общие тренды активности обсуждений соответствуют картине, предоставляемой специализированными сервисами аналитики. В анализе общих чатов присутствует человеческий фактор, что позволяет отсечь неинтересные темы для обсуждения, завышенные показатели из-за деятельности центров анализа трендов и темы разрекламированные новостями, как это видно на примере с Google Trends, поскольку завышенная история запросов не позволяет увидеть локальные всплески.

Перспективы развития

Данная работа потребовала обширного технического погружения в реализацию. К данным трудностям можно отнести проблемы с доступом к серверам, настройку CI/CD составляющей и неопределённость в выборе первоначальных коэффициентов анализа.

В перспективе планируется выявить и реализовать наиболее точные методы аналитики. В частности уделить особое внимание методу BTM, предназначенному для анализа малых фрагментах данных, подобно текстовым сообщениям в мессенджерах.

В целях исключения таких явлений, как реклама, «вбросы» и сообщения, не являющиеся форматом естественного общения, необходимо реализовать алгоритм фильтрации контента. Также необходимо отсеивать повторяющиеся сообщения.

В настоящих условиях возникла сложность с реализацией графа взаимосвязи объектов дискуссии и пользователей между собой. Причиной стало неравномерное распределение сообщений большого и малого объёма. Т.е. ветвь связей зачастую прерывалась.

Интересным для анализа является алгоритм формирования биграмм и триграмм (шинглов), так как в сообщениях зачастую встречаются устойчивые выражения, названия и сокращения.

Добавление в анализ большего количества информации о сообщениях и пользователях позволит учитывать перекрёстных пользователей в различных чатах, а также географическую причастность.

В текущей работе не предусмотрена возможность изменения диапазона анализа и временных когорт. В частности это позволит отображать график активности обсуждений по заданным дискуссиям в заданный промежуток времени.

Список литературы

[1] Top Websites Ranking // SimilarWeb [2020]. Дата обновления: 01.04.2020. URL: https://www.similarweb.com/top-websites (дата обращения: 28.05.2020)

[2] Lopez-Castroman J, Moulahi B, Azé J, et al. «Mining social networks to improve suicide prevention: A scoping review». J Neurosci Res. 2020; 98:616–625.

[3] Facebook Reports First Quarter 2020 Results // Facebook [2019]. URL: https://s21.q4cdn.com/399680738/files/doc_financials/2020/q1/Q1-2020-FB-Earnings-Presentation.pdf (дата обращения: 27.05.2020)

[4] Global Digital Report 2019 // We Are Social [2008-2020]. URL: https://wearesocial.com/global-digital-report-2019 (дата обращения: 28.05.2020)

[5] Екатерина Курносова, Социальные сети в цифрах // Российский интернет-форум | РИФ+КИБ, 2019. URL: https://mediascope.net/upload/iblock/f97/18.04.2019_Mediascope_Екатерина Курносова_РИФ+КИБ 2019.pdf (дата обращения 28.05.2020)

[6] Number of monthly active WeChat users from 2nd quarter 2011 to 1st quarter 2020 // Statista [2020]. Дата обновления: 05.2020. URL: https://www.statista.com/statistics/255778/number-of-active-wechat-messenger-accounts (дата обращения: 28.05.2020)

[7] Telegram Reports // Telegram [2020]. Дата обновления: 24.04.2020. URL: https://telegram.org/blog/400-million?ln=f (дата обращения 28.05.2020)

[8] 15 Billion Telegrams Delivered Daily // Telegram [2020]. Дата обновления: 23.02.2016. URL: https://telegram.org/blog/15-billion (дата обращения: 29.05.2020)

[9] About // YouScan [2020]. URL: https://youscan.io/ru/about/ (дата обращения: 29.05.2020)

[10] Нигматуллина К.Р., Бодрунова С.С. Методика качественного анализа дискуссий в Twitter // Медиаскоп. 2017. Вып. 1. Режим доступа: http://www.mediascope.ru/2293

[11] Садреева Ю. И., Добрынин В. Ю. Выпускная квалификационная работа бакалавра «Автоматическая классификация новостей из коллекции Reuters в таксономию IPTC» // Архив открытого доступа Санкт-Петербургского государственного университета [2016]. URL: https://dspace.spbu.ru/bitstream/11701/4113/1/VKR_Sadreeva.pdf (дата обращения: 20.12.2019)

[12] S. S. Bodrunova, I. S. Blekanov and M. Kukarkin, "Topics in the Russian Twitter and Relations between their Interpretability and Sentiment,"2019 Sixth International Conference on Social Networks Analysis, Management and Security (SNAMS), Granada, Spain, 2019, pp. 549-554, doi: 10.1109/SNAMS.2019.8931725.

[13] К. В. Воронцов, Вероятностное тематическое моделирование // 16.10.2013. URL: http://www.machinelearning.ru/wiki/images/2/22/Voron-2013-ptm.pdf

[14] Пархоменко П.А., Григорьев А.А., Астраханцев Н.А. Обзор и экспериментальное сравнение методов кластеризации текстов. // Труды ИСП РАН, 2017 г., том 29, вып. 2, с. 161-200 URL: https://www.ispras.ru/proceedings/docs/2017/29/2/isp_29_2017_2_161.pdf

[15] Girolami, Mark; Kaban, A. (2003). On an Equivalence between PLSI and LDA. Proceedings of SIGIR 2003. New York: Association for Computing Machinery. ISBN 1-58113-646-3.